ОЦЕНКИ ВАРИОГРАММЫ СТАЦИОНАРНОГО СЛУЧАЙНОГО ПРОЦЕССА

Рассмотрим внутренне стационарный с независимыми приращениями случайный процесс $X(s), s \in Z$. Предположим, что $MX(s) = 0, s \in Z$, и неизвестна вариограмма

$$2\gamma(h) = D(X(s+h) - X(s)), s, h \in Z.$$

Вариограмма является одной из важнейших статистических характеристик при анализе временных рядов в экономике, финансах, геологии, экологии, ... Однако проблема построения ее оценки значительно упрощается в научной литературе. Это связано с тем, что обычно делается предположение о нормальном распределении исходных данных [1, 2]. На практике обнаруживается, что данные не подчинены нормальному закону, и условие стационарности в широком смысле является достаточно сильным. В данной работе обобщаются вопросы исследования оценок вариограммы, рассмотренные в [3], для внутренне стационарных случайных процессов.

Обозначим $U_s(h) = (X(s+h) - X(s))^2, s, h \in Z$. Тогда $2\gamma(h) = MU_s(h)$, то есть оценивание вариограммы сводится к построению оценки математического ожидания процесса $U_s(p), p, s \in Z$. Пусть $X(1), X(2), \dots, X(n)$ – n последовательных, полученных через равные промежутки времени наблюдений за процессом $X(s), s \in Z$, и $U_s(h), h = \overline{1, n-1}, s = \overline{1, n-h}$, – одинаково распределенные величины с эмпирической функцией распределения $F_{n-h} = F_{n-h}(u)$. Следовательно, из определения величин $U_s(h), h = \overline{1, n-1}, s = \overline{1, n-h}$, вытекает их независимость [4].

Рассмотрим оценки для $2\gamma(h)$ вида:

1. Выборочное среднее

$$\bar{U}(h) = \frac{1}{n-h} \sum_{s=1}^{n-h} U_s(h),$$

$$h = \overline{1, n-1};$$

2. Выборочная медиана

$$\tilde{U}(h) = \begin{cases} \frac{\bar{U}_{\frac{n-h+1}{2}}(h), & (n-h) \text{ нечетно,} \\ 0,5 \left(\frac{\bar{U}_{\frac{n-h}{2}}(h) + \bar{U}_{\frac{n-h+1}{2}}(h)}{2} \right), & (n-h) \text{ четно,} \end{cases}$$

где $\bar{U}_1(h) \leq \bar{U}_2(h) \leq \dots \leq \bar{U}_{n-h}(h)$ – порядковые статистики для $U_s(h), h = \overline{1, n-1}, s = \overline{1, n-h}$;

3. Урезанное среднее

$$T(\alpha) = \sum_{j=k+1}^{n-h-k} \frac{\bar{U}_j(h)}{n-h-2k},$$

где $\bar{U}_j(h), j = \overline{1, n-h}$, – порядковые статистики для $U_s(h), h = \overline{1, n-1}, s = \overline{1, n-h}, k = [(n-h)\alpha],$

$\alpha \in [0, \frac{1}{2}]$. Эту оценку получают исключением k крайних порядковых статистик $U_j(h)$ и усреднением оставшихся. Отметим, что выборочное среднее и выборочная медиана – два случая симметричного урезания, когда либо не отбрасывается ни одного наблюдения, т.е. $\alpha = 0, T(\alpha) = \bar{U}(h)$, либо отбрасываются все, за исключением одного или двух наблюдений в центре выборки.

В работе рассматриваются следующие параметры урезания:

- 3.1. $\alpha = 0,001$;
- 3.2. $\alpha = 0,05$;
- 3.3. $\alpha = 0,1$;
- 3.4. $\alpha = 0,25$;

4. M -оценки T_n , которые удовлетворяют соотношению

$$\inf_t \sum_{j=1}^{n-h} \eta(U_j(h) - t) = \sum_{j=1}^{n-h} \eta(U_j(h) - T_n),$$

где $\eta(\cdot)$ – некоторая действительная функция [5]. Например, $\eta(x) = x^2$, $x \in R$, и $\eta(x) = |x|$, $x \in R$, соответствуют выборочному среднему и выборочной медиане. Предположим, что $\eta(x)$ – дифференцируемая функция и $\eta'(x) = \varphi(x)$, тогда оценка T_n является решением уравнения

$$\sum_{j=1}^{n-h} \varphi(U_j(h) - T_n) = 0.$$

Исследуются оценки вида:

4.1. M -оценка Хьюбера с

$$\varphi(x) = \varphi_1(x) = \begin{cases} x, & |x| \leq k, \\ \text{sgn}(x), & |x| > k, \end{cases} \quad k = 1,4088;$$

4.2. M -оценка Тьюки с

$$\varphi(x) = \varphi_2(x) = \begin{cases} x(r^2 - x^2)^2, & |x| \leq r, \\ 0, & |x| > r, \end{cases} \quad r = 4;$$

4.3. M -оценка Хампеля с

$$\varphi(x) = \varphi_3(x) = \begin{cases} x, & 0 \leq |x| \leq a, \\ a \text{sgn}(x), & a \leq |x| \leq b, \\ a \frac{r - |x|}{r - b} \text{sgn}(x), & b \leq |x| \leq r, \\ 0, & |x| \geq r, \end{cases} \quad a = 1,31, \quad b = 2,039, \quad r = 4;$$

4.4. M -оценка Эндрюса с

$$\varphi(x) = \varphi_4(x) = \begin{cases} \sin(x), & |x| \leq pa, \\ 0, & |x| > pa, \end{cases} \quad a = 1,142.$$

При исследовании M -оценок каждому наблюдению приписывается вес, определяемый функцией $\varphi(x)$, и значения резко выделяющихся наблюдений можно сгладить с помощью этой функции.

Семейство рассматриваемых оценок 1 – 3 можно представить в виде

$$T_{n-h} = t_{n-h}(F_{n-h}) = t_{n-h}(\overline{U}_1(h), \overline{U}_2(h), \dots, \overline{U}_{n-h}(h)),$$

$h = \overline{1, n-1}$. Следовательно, оценки математического ожидания процесса $U_s(p)$, $p, s \in Z$, получаются посредством функциональных преобразований порядковых статистик $\overline{U}_1(h), \overline{U}_2(h), \dots, \overline{U}_{n-h}(h)$. Для среднего, например, функциональная форма не зависит от n и $\overline{U}(h) = t(F_{n-h})$. В других

случаях можно выбрать подходящий функционал $t(F_{n-h})$ так, что при больших n разность $t_n(F_{n-h}) - t(F_{n-h})$ будет сколь угодно малой. Если оценка вариограммы определяется выборочной медианой, то функционал $t(F_{n-h}) = F_{n-h}^{-1}(1/2)$, а если задается урезанным средним, то функционал имеет вид

$$t(F_{n-h}) = \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F_{n-h}^{-1}(v) dv.$$

Для M -оценок $t(F_{n-h})$ неявно определяется уравнением вида

$$\int \varphi(u - t(F_{n-h})) dF_{n-h}(u) = 0.$$

Известно, что если $t(\cdot)$ – непрерывный функционал, то оценка $t(F_{n-h})$ состоятельна для $t(F)$, где $F = F(u)$ – истинное распределение $U_s(h)$, $h = \overline{1, n-1}$, $s = \overline{1, n-h}$. Далее, следуя работам [5, 6], можно показать, что статистика

$$\frac{\{T_{n-h} - t(F)\} \sqrt{n-h}}{y_{t,F}}$$

имеет асимптотически нормальное распределение с нулевым математическим ожиданием и дисперсией, равной единице, где

$$y_{t,F}^2 = \int d_{t,F}^2(u) dF(u), \quad (1)$$

$d_{t,F}(u)$ – производная фон Мизеса, которая определяется как

$$\lim_{\varepsilon \rightarrow 0} \frac{t((1-e)F + eK) - t(F)}{e} = \int d_{t,F}(z) dK(z). \quad (2)$$

Производную $d_{t,F}(u)$ можно получить из (2), полагая

$$K(z) = hv(z-y) = \begin{cases} 1, & z \geq y, \\ 0, & z < y, \end{cases}$$

где $hv(z)$ – единичная функция Хевисайда.

Так, для урезанного среднего $T(\alpha)$, $0 \leq \alpha < \frac{1}{2}$, производная фон Мизеса имеет вид

$$d_{t,F}(u) = \begin{cases} \frac{o_\alpha}{(1-2\alpha)}, & u < o_\alpha, \\ \frac{u}{(1-2\alpha)}, & o_\alpha \leq u \leq o_\alpha, \\ \frac{o_{1-\alpha}}{(1-2\alpha)}, & o_{1-\alpha} < u, \end{cases}$$

где $o_\alpha = F^{-1}(\alpha)$, а дисперсия

$$y_{t,F}^2 = \frac{1}{(1-2\alpha)^2} \left(\int_{\xi_\alpha}^{\xi_{1-\alpha}} u^2 dF(u) + 2\alpha \xi_\alpha^2 \right).$$

Для выборочной медианы

$$d_{t,F}(u) = \frac{\text{sgn}(u)}{2f(0)},$$

$$y_{t,F}^2 = \frac{1}{(2f(0))^2},$$

где $f(u)$ – плотность распределения, соответствующая функции распределения $F(u)$. В случае M - оценок

$$d_{t,F}(y) = \frac{\varphi(y)}{\int \varphi'(u)f(u)du},$$

$$\sigma_{t,F}^2 = \frac{\int \varphi^2(y)f(y)dy}{\left[\int \varphi'(u)f(u)du\right]^2}.$$

Производная фон Мизеса $d_{t,F}(u)$ является удобным инструментом измерения влияния резко выделяющихся наблюдений на оценки вариограммы. Это связано с тем, что на ней основана одна из важнейших характеристик – чувствительность к большой ошибке, определяемая соотношением

$$\gamma^* = \sup_u |d_{t,F}(u)|,$$

Таким образом, наибольшее влияние на значение оценки загрязнения фиксированного объема характеризуется чувствительностью γ^* .

Чтобы проиллюстрировать свойства вышеперечисленных оценок 1 – 4, приведем некоторые численные результаты, относящиеся к семи различным порождающим симметричным распределениям, а именно

- A. Стандартное нормальное распределение;
- B. Смесь нормальных распределений $N(0,1)$ и $N(0,9)$ с весами $\varepsilon = 0,95$ и $(1 - \varepsilon) = 0,05$ соответственно;
- C. Смесь нормальных распределений $N(0,1)$ и $N(0,9)$ с весами $\varepsilon = 0,9$ и $(1 - \varepsilon) = 0,1$ соответственно;
- D. Смесь нормальных распределений $N(0,1)$ и $N(0,9)$ с весами $\varepsilon = 0,75$ и $(1 - \varepsilon) = 0,25$ соответственно;
- E. Смесь нормальных распределений $N(0,1)$ и $N(0,100)$ с весами $\varepsilon = 0,95$ и $(1 - \varepsilon) = 0,05$ соответственно;
- F. Распределение Лапласа с плотностью

$$f(x) = \frac{e^{-|x|}}{2}, \quad x \in R;$$

- G. Распределение Коши с плотностью

$$f(x) = \frac{1}{p(1+x^2)}, \quad x \in R.$$

В случаях B – D функция распределения имеет вид:

$$F(u) = e \Phi(u) + (1 - e) \Phi\left(\frac{u}{3}\right),$$

в случае E –

$$F(u) = e \Phi(u) + (1 - e) \Phi\left(\frac{u}{10}\right),$$

где $\Phi(u)$ – функция распределения стандартного нормального закона.

Для каждой комбинации оценки и порождающего распределения в таблице найдены значения дисперсии $y_{t,F}^2$ предельного нормального распределения статистики $\sqrt{n-h}(T_{n-h} - t(F))$, вычисленные по формуле (1),

Таблица

Асимптотическая дисперсия статистики $\sqrt{n-h}(T_{n-h} - t(F))$
 для десяти оценок T_{n-h} ,
 определенных в пунктах 1 – 4, и семи порождающих распределений.

Оценки	Распределения						
	А	В	С	Д	Е	Ф	Г
1	1,0000	1,4000	1,8000	3,0000	5,9500	2,0000	∞
2	1,5708	1,6810	1,8032	2,2620	1,7223	1,0000	2,4674
3.1	1,0003	1,3739	1,7749	2,9791	5,6217	1,9791	304,69
3.2	1,0263	1,1554	1,3173	2,1378	1,2462	1,6537	6,8470
3.3	1,0604	1,1688	1,2964	1,8490	1,2276	1,4941	3,8658
3.4	1,1952	1,2892	1,3949	1,8039	1,3285	1,2274	2,2732
4.1	1,0457	1,1649	1,2968	1,7877	1,2273	1,4406	2,9033
4.2	1,0989	1,1978	1,3107	1,7645	1,1780	1,4077	2,2593
4.3	1,0966	1,1954	1,3080	1,7603	1,1757	1,4558	2,3000
4.4	1,0998	1,1991	1,3125	1,7687	1,1789	1,4133	2,2687

Полученные численные результаты позволяют выявить наиболее устойчивые оценки вариограммы $2\gamma(h)$.

Как видно из таблицы, выборочное среднее приемлемо при стандартном нормальном порождающем распределении, но совершенно не пригодно, где возможно появление резко выделяющихся наблюдений. Это объясняется тем, что чувствительность к большой ошибке выборочного среднего $\gamma^* = +\infty$. Следовательно, даже одно резко выделяющееся наблюдение может оказать сильное влияние на значение оценки вариограммы. Значение асимптотической дисперсии выборочного среднего для распределения Коши можно объяснить поведением хвостов плотности этого распределения.

Резко выделяющиеся наблюдения оказывают влияние и на выборочную медиану, которая обладает конечной чувствительностью $\gamma^* = 1,253$. Заметим, что выборочная медиана оказывается предпочтительнее выборочного среднего для распределений с "тяжелыми хвостами".

В качестве альтернативы между выборочным средним и выборочной медианой выступает урезанное среднее, которое дает возможность получения более устойчивой оценки вариограммы путем отбрасывания $[(n-h)\alpha]$ наибольших и $[(n-h)\alpha]$ наименьших наблюдений. Это имеет значение при наличии загрязнения резко выделяющимися наблюдениями. Применение урезанного среднего требует предварительной оценки оптимальной доли урезания α для каждого порождающего распределения. Параметр α считается оптимальным для порождающего распределения F , если он минимизирует дисперсию $y_{i,F}^2$, заданную равенством (1).

Из полученных результатов следует, что M - оценки являются наиболее стабильными для всех распределений. Это объясняется их способностью полного удаления резко выделяющихся наблюдений. M - оценки Хампеля, Тьюки и Эндрюса ведут себя похожим образом. M - оценка Хьюбера остается лучшей при распределениях с относительно небольшим загрязнением (случаи В, С). Но при распределениях с большой долей резко выделяющихся наблюдений, ее дисперсия существенно возрастает.

Описанный выше метод оценивания вариограммы случайного процесса является дальнейшим развитием некоторых результатов, приведенных в монографии Д. Кокса, Д. Хинкли [6]. В частности, ими исследованы пять оценок параметра положения, а именно: выборочные среднее и медиана, урезанное среднее с оцениваемой оптимальной долей урезания α , M – оценка

Хьюбера при $k = 1$, линейная комбинация трех порядковых статистик и рассмотрены четыре порождающих распределения, определенных в пунктах А, С, F, G.

В случае если процесс $X(s)$, $s \in Z$, является процессом стационарным в широком смысле, то вариограмма $2\gamma(h)$, $h \in Z$, связана с ковариационной функцией $R(h)$, $h \in Z$, следующим соотношением:

$$\gamma(h) = \{R(0) - R(h)\}, h \in Z.$$

В статистическом анализе временных рядов оценкам ковариационной функции уделяется большое внимание. Исследованию моментов оценки ковариационной функции и их асимптотическому поведению посвящены работы [7, 8].

В качестве оценки вариограммы можно рассматривать статистику вида:

$$2\tilde{\gamma}(h) = \frac{1}{n-h} \sum_{s=1}^{n-h} (Y(s+h) - Y(s))^2, \quad (3)$$

$h = \overline{0, n-1}$, $\tilde{\gamma}(-h) = \tilde{\gamma}(h)$, $h = \overline{0, n-1}$ и $\tilde{\gamma}(h) = 0$ для $|h| \geq n$.

В работе [7] было показано, что оценка (3) является несмещенной и состоятельной в среднеквадратическом смысле, а в статье [9] найдено ее предельное распределение.

Summary

The variogram estimators of the intrinsic stationary stochastic process with discrete time are constructed. The statistical properties of the examined estimators are investigated.

Литература

1. Davis B. M., Borgman L. E. // *Mathematical Geology*. 1979. Vol. 11, No. 6. P. 643 – 653.
2. Cressie N. // *Mathematical Geology*. 1985. Vol. 17, No. 5. P. 563 – 586.
3. Cressie N., Hawkins D. M. // *Mathematical Geology*. 1980. Vol. 12, No. 2. P. 115 – 125.
4. Боровков А.А. Теория вероятностей. М., 1986.
5. Huber P.J. // *The Annals of Mathematical Statistics*. 1964. Vol. 35. P. 73–101.
6. Кокс Д., Хинкли Д. Теоретическая статистика. М., 1978.
7. Труш Н.Н., Цеховая Т.В. // *Весті НАН Беларусі. Сер. фіз.-мат. навук*. 2001. №2. С. 24–29.
8. Цеховая Т.В. // *БелИСА*, № Д20011 от 26.01.2001. Выпуск 1 (20), 2001.
9. Цеховая Т.В. // *Вестник БГУ, серия физ.-мат.* 2002. №1. С. 104 – 105.